MONTE CARLO SIMULATION AND STATISTICAL ANALYSIS OF GENETIC INFORMATION CODING

E. Gultepe * M. L. Kurnaz **

Department of Physics, Bogazici University, 34342 Bebek Istanbul

Abstract

The rules that specify how the information contained in DNA codes amino acids, is called "the genetic code". Using a simplified version of the Penna nodel, we are using computer simulations to investigate the importance of the genetic code and the number of amino acids in Nature on population dynamics. We find that the genetic code is not a random pairing of codons to amino acids and the number of amino acids in Nature is an optimum under mutations.

1 INTRODUCTION

In general population dynamics is a matter of interest for biologists, however it has attracted the attention of physicists since it is a subject very closely related to statistical mechanics. Investigation of population dynamics in Nature is not a simple task because to get any idea about the dynamics of population growth, one has to consider many generations of a population. Even if one can find fast-reproducing species like the fruit fly, checking all individuals in such a population is not an easy task either. Therefore, modelling with computers has lots of advantages such as considerably small time consumption and simplicity in population monitoring.

The most successful computational model for age-structured populations is the Penna model [1]. In this model, individuals are represented by bit-strings which are 32 bits long and are initially set to zero. Each bit represents a given

* Present Address: Northeastern University
**Corresponding Author
Email address: kurnaz@boun.edu.tr (M. L. Kurnaz).

age: as the individual gets older we move down on the bit-string. Bits which are set to zero represent that no bad mutations is stored at that age. However, if a bit is set to one, it means that the individual suffers a disease at that age and its probability of staying alive is decreased.

The Penna model has been successfully used to investigate the advantages of sexual reproduction over asexual reproduction [2][3][4][5][6][7][8], certain features of ecology [9] and population dynamics [10][11][12].

To investigate the importance of *the genetic code* and number of amino acids in population dynamics we have constructed a model based on the Penna model.

The genetic information about the individuals is stored in the DNA. DNA is made up of different monomers. Each monomer, nucleotide, in the chain carries a heterocyclic base. In DNA, these bases are adenine (A), guanine (G), cytosine (C) and thymine (T). Proteins are synthesized from amino acids using the information stored in the DNA. As there are four bases in the DNA, and 20 amino acids used in proteins, during protein synthesis there is not a one-to-one correspondence between the nucleotides in the DNA and the amino acids in a protein. Rather, the linear sequence of bases which constitutes the protein-coding information is "read" by the cell in blocks of three nucleotide residues, or codons, each of which specifies a different amino acid. If we consider a nucleotide on the DNA to be a letter in a four-letter alphabet, codons can be thought as words with three letters. Hence, there are sixty four words to code the twenty different amino acids. The set of rules that specifies which nucleic acid codons correspond to which amino acid is known as the genetic code.

2 COMPUTATIONAL METHOD

If there is a mutation on a gene which causes a change in the amino acid chain, we will think that the organism may not be able to build the protein which may be crucial for the organism. If so; it will not function properly or it may simply die; hence it is simplistic yet reasonable to represent the organism by a single gene.

In our model, to represent a whole individual, we took a real gene from Nature "human cytokine" (LD78 Homo sapiens blood lymphocyte gene on the DNA 17^{th} chromosome) [13]. This gene is necessary for activating lymphocytes; therefore if it is missing the human body cannot perform immune responses.

If all other effects (aging, food restriction, illness etc.) are neglected, muta-

tion will be the only possible cause of death. Also, in our simplistic model reproduction is not included, therefore we have a population which can only decrease as a result of mutations. We use this model to investigate the effects of mutations to population decrease.

A mutation in our model is a process acting at each site independently. We disregard more complicated processes such as deletions or insertions, and we only look at single nucleotide replacements by another nucleotide in the gene. Normally the rates for these replacements depend on the two nucleotides being interchanged. The simplest approach to the problem is to take all mutation rates to be equal, known as the Jukes-Cantor mutation scheme [14].

The mutation is taken to be deleterious if it causes a change in the amino acid chain; and not all the mutations kill the individual. A real gene is composed of two different parts: a coding portion and a noncoding portion. The coding part, exon, is responsible for coding for proteins whereas the rest, intron, does not code for a protein and the purpose of this part is not clearly understood yet. If a mutation takes place on intron part, it is considered to be simply harmless but if it takes place on exon part, it is usually harmful, but there is still a chance: The interchanged codon may still code the same amino acid since more than one codon can code one amino acid in nature.

To be more explicit, the codons AAA and AAG code the same amino acid, "lysine"; hence if AAA turns into AAG as a result of a mutation the amino acid will not change and the protein can be constructed safely. However; if AAA turns into AGA, which codes the amino acid "arginine", the amino acid chain will change and we assume that the protein can not build up, which means the represented organism will die.

There can be a mutation which converts AAA to AAX where X $\neq$ A, G, C, T; then the individual dies automatically. As a model, we are looking at a simpler case where a mutation changes A to one of G, C, T not X.

Since reproduction is not included in the model, the population can only diminish. The decrease in population can be found by calculating the probability of a deleterious mutation. The probability of the mutation changing the amino acid depends on the codon; so one needs to find the probability of hitting each different codon type. First, the probability of hiting a codon type (P_α) is calculated as the ratio of the number of codons of that type in the gene (N_α) to total number of codons. Then we need to exclude the mutations that do not cause a change in the amino acid and calculate the probability of a change occurring in the amino acid caused by a change in one nucleotide ($P(d/\alpha)$) is calculated.

As an example; only two codons code the amino acid "lysine": AAA and AAG. In the exon part of human cytokine gene, there are only three "AAA" codons

and the total number of codons in the gene is 207, hence the probability of the mutation hitting an "AAA" codon is simply $P_\alpha = 3 \div 207 = 0.0145$. By a point mutation to "AAA" we can have 9 different codons (AAC, AAG, AAU, ACA, AGA, AUA, CAA, GAA, UAA). One of these codons still codes the same amino acid (AAG). Therefore the probability of deleterious mutation ($P(d/\alpha)$) is 8/9 for "AAA" in the human cytokine gene.

Next, we need to calculate the probability of hitting the exon part of the gene as the ratio of the exon part to the total gene. In the human cytokine gene, there are 621 nucleotides in exon part and 1447 ones in intron part:

$$P(hitting\, exon) = \frac{621}{2068} = 0.3032 \tag{1}$$

Hence; the probability of having a deleterious mutation for all of the gene is simply:

$$P(deleterious) = P(hitting\, exon) \sum_{\alpha=1}^{64} [P_\alpha P(d/\alpha)] = 0.2344 \tag{2}$$

The survival probability can be calculated by:

$$P(surviving) = 1 - P(deleterious) = 0.7656 \tag{3}$$

If we take a population of N_0 gene (individuals), after n mutations, to the first approximation, the number of surviving individuals is given by:

$$N_n \approx N_0 P(surviving)^n \tag{4}$$

Hence, we obtain the "probability of survival" with the slope of the number of surviving individuals versus time graph:

$$slope \approx ln[P(surviving)] = -0.2670 \tag{5}$$

During this calculation we used a simple assumption that after each timestep the genome remain the same as the wildtype. A mutation may result in a different nucleotide sequence, but if this sequence codes the same amino acid, we assume that this mutation has never happened and go to the next stage. Hence, all alive individuals can be represented by the same array, wildtype, as in the calculations. However, in the less likely event of a harmless mutation the number of the codons of each type changes which will in turn slightly change the probabilities. We have designed a test simulation where after each mutation and deletion of the individual, we have set all the sequence back to

the wildtype. As this simulation gives the same results (within the error bars) as the original case, we have used the modified sequence in the later stages.

3 SIMULATION

In the simulation, an individual (a gene) is represented by an array which contains 0, 1, 2, and 3's instead of the nucleotides Adenine (A), Guanine (G), Cytosine (C) and Thymine (Uracil (U)) respectively and also a sign bit which shows if the gene has a deleterious mutation (1) or not (0).

In each "cycle", each individual has to go through one mutation event, then it is determined whether or not the individual should die. In the mutation event; the place of mutation and the mutant nucleotide is determined randomly. If the nucleotide is not in the exon part, the sign bit remains 0. Otherwise, the changed codon is checked for the amino acid which it codes. If it is coding the same amino acid, the protein can still be built, therefore the sign bit is not changed and the individual survives. However, if the amino acid is changed then the sign bit becomes '1' that means this individual will be deleted. Deletion time is recorded for each individual [Fig.1 (a)]. In the control simulation, if the mutation is harmless, modification of the gene will be recovered [Fig.1(b)].

After N individuals, the number of surviving individuals in each time step is calculated. Since the probability of mutation is independent of the number of individuals, this number also gives us the population size. Hence, we have an exponential population decay and the exponent depends on the probability of surviving ($P(surviving)$). Logarithm of the population is fitted to a straight line and the slope of the line is calculated.

We run all simulations ten times. The average of the slopes of the control simulations -0.266 ± 0.001, which is noticeably close to the slope derived from calculations. After the control, we run the simulation using genetic code of Nature. One example of such runs is shown in Figure 2. The average of slopes for Nature's simulation is -0.266 ± 0.001.

4 ARTIFICIAL TABLES

With a few exceptions, twenty different kinds of amino acids are used to build the proteins. Even though in some rare cases certain organisms use selenocys-

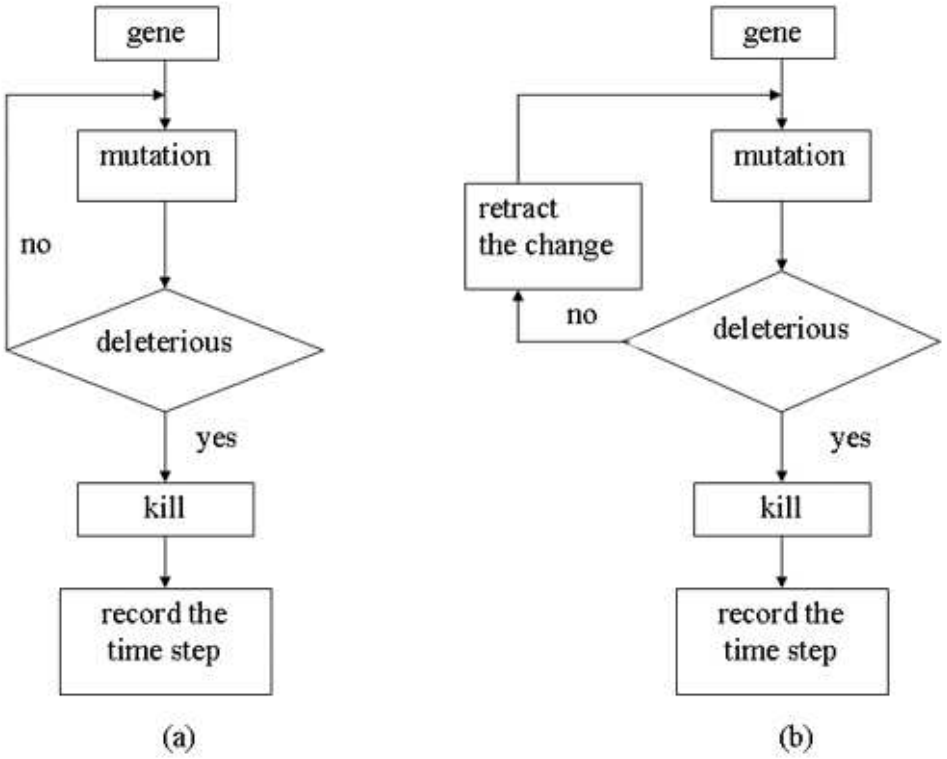

Fig. 1. a.) Flowchart of the simulation b.) Flowchart of the control simulation

teine and pyrrolysine, in Nature, the majority uses the same table. Recently a team of investigators at the Scripps Research Institute modified a form of the bacterium Escherichia coli to use a 22-amino acid genetic code instead of 20 [15]. They have engineered the modified form of E. Coli to make myoglobin proteins with 22 amino acids, using the unnatural amino acids O-methyl-L-tyrosine and L-homoglutamine in addition to the naturally occurring 20. This work opens up the possibility that the same procedure can be used to expand the amino acid family even further. So, the question is why did life stop with twenty amino acids? To investigate the importance of the number of amino acids, we create amino acid tables based on Nature's table but with different amino acid numbers and we use them in the simulation.

If we change the number of amino acids in the genetic code, it means that we change the amount of information in the genome. Hence, to conserve the information, the genome length needs to be adjusted also. Moreover, if we want our simulation to represent Nature, the process of extending or shortening the amino acid table needs to obey some rules of biochemistry.

The twenty amino acids contain with their twenty different side chains of different chemical properties. This allows proteins to have such a great variety of structures and properties. There are several classes of side chains, grouped

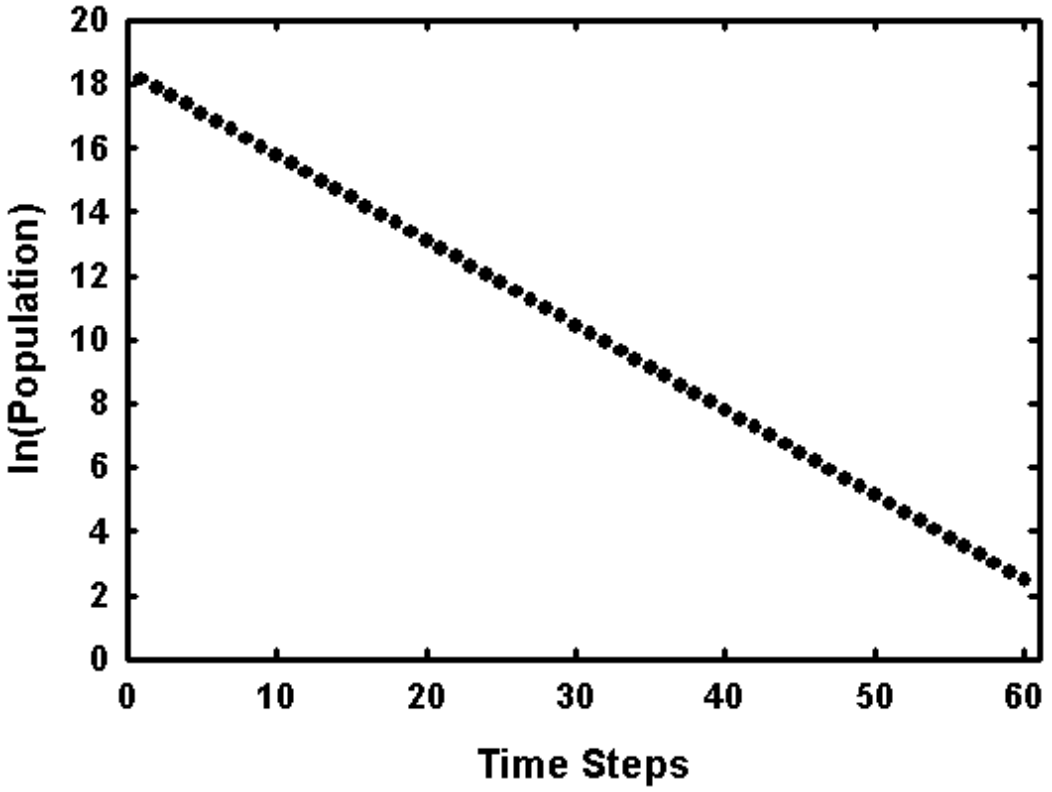

Fig. 2. Population decreasing: one of the simulations of the amino acids table of Nature

by their dominant chemical features. While developing tables, we try to make them fit the natural structure of amino acids obeying the classification in [16].

We have tried to use amino acid tables with more and with less number of amino acids. To shorten the amino acid tables, we first randomly choose which amino acid will be removed from the table. The random choice is made such that no two amino acids are removed from the same group (as long as the number of removed amino acids is less than the number of amino acids groups).

For example, in the table which has 18 amino acids Glutamine and Isoleucine are removed. Glutamine is in the acidic group and its frequency of occurrence is 3.9%. The codons which code Glutamine formerly (CAA, CAG) will code Glutamic Acid which is also in acidic group and has the frequency of 6.2%. Isoleucine is in the aliphatic group and its frequency is 4.6%. The codons which code Isoleucine formerly (AUU, AUC, AUA) will code Glycine which is also in aliphatic group and has the frequency of 7.5%.

To conserve the information content of the gene, the gene should be lengthened. As an example, by changing the codons CAA and CAG from Glutamine to Glutamic Acid we have lost the information carried by Glutamine. Now we take Asparagine and Aspartic Acid, which are also in the acidic group, and insert them where Glutamine was originally. The same procedure was repeated

to decrease the number of amino acids to 16 and then to 14.

To extend the amino acid table, first we determine which consecutive amino acid pair in the gene has the highest frequency. To do this, we calculate the number of occurrence of pairs and construct the pair matrix. Then, each frequent pair is replaced by a new amino acid. For example, Leucine - Leucine pair has the highest frequency and they will be replaced by the new amino acid called New1. Similarly Threonine - Serine pairs are replaced by New2. Leucine is from the aliphatic group, and New1 is constructed by dividing Alanine, which is also from the aliphatics group and is represented by the highest number of codons. Now the codons GCU and GCC still code Alanine but the remaining (GCA, GCG) will code New1. Similary, New2 is formed by dividing Serine: the codons UCU, UCC, UCA, and UCG code still Serine but the others (AGU and AGG) code New2. The tables with 24 amino acid , with 26 amino acids and with 28 amino acids are constructed just the same way.

As control cases we have also done simulations with different amino acid tables, both increased and decreased number of amino acids, where we neglected the conservation of information and kept the genome length constant. As expected, the results of these simulations were very close (within error bars) to the results of the original table.

Biologists have also been trying to find simplified amino acid alphabets. One of these methods is the MJ matrix constructed using Wang and Wang's method [17] which is based on Miyazawa-Jernigan's (MJ) residue - residue potentials [18]. Their reduction algorithm, which connects different representations of a protein, is generally based on the idea that the amino acids can be distributed into different groups, with different interactions. The interactions between amino acids of two different groups should have similar characteristics for a successful reduction.

Another method is the BLOSUM50 matrix, built using Murphy, Wallqvist and Levy's method [19] derived by Henikoff and Henikoff [20]. Their reduction scheme is based on the analysis of correlations among similarity matrix elements used for sequence alignment. We have constructed reduced amino acid tables using both the MJ matrix and BLOSUM50 matrix methods.

5 CONCLUSION

In this paper, we developed a computer simulation which represents a living organism under mutations. Furthermore, we changed the genetic code used in the simulations to analyze its effect on population stability.

Table 1
Results of the simulations using different genetic code tables.

Table Name	"probability of survival"
Nature	-0.266 ± 0.001
with 18	-0.281 ± 0.001
with 16	-0.291 ± 0.004
with 14	-0.313 ± 0.004
with 18 (using MJ Matrix)	-0.282 ± 0.002
with 16 (using MJ Matrix)	-0.287 ± 0.003
with 14 (using MJ Matrix)	-0.320 ± 0.003
with 18 (using BLOSUM50 Matrix)	-0.288 ± 0.003
with 16 (using BLOSUM50 Matrix)	-0.294 ± 0.003
with 14 (using BLOSUM50 Matrix)	-0.302 ± 0.004
with 22	-0.265 ± 0.001
with 24	-0.265 ± 0.001
with 26	-0.273 ± 0.001
with 28	-0.291 ± 0.002

All the results of different simulations are summarized in Table 1 and plotted in Figure 3.

The results of shorter amino acid tables show that if we try to conserve the information, the population which is represented by less amino acids is affected more severely by mutations. However, if we do not mind the information transferred by the gene (the simulations with conserved genome length), the population is not affected much.

Even if we use different reducing algorithms (MJ or BLOSUM50) for genetic code, the population is affected by mutations more than the population represented by the genetic code of Nature.

While we extend the amino acid table, we shorten the size of the gene which means that we try to conserve the information. The resistance of the population against mutations does not change when the amino acid number is 22 or 24. However, after 24, the resistance starts to decrease.

The slopes of the simulations with 20, 22 and 24 amino acids are very close. These results are along the same line with the results of the investigators from

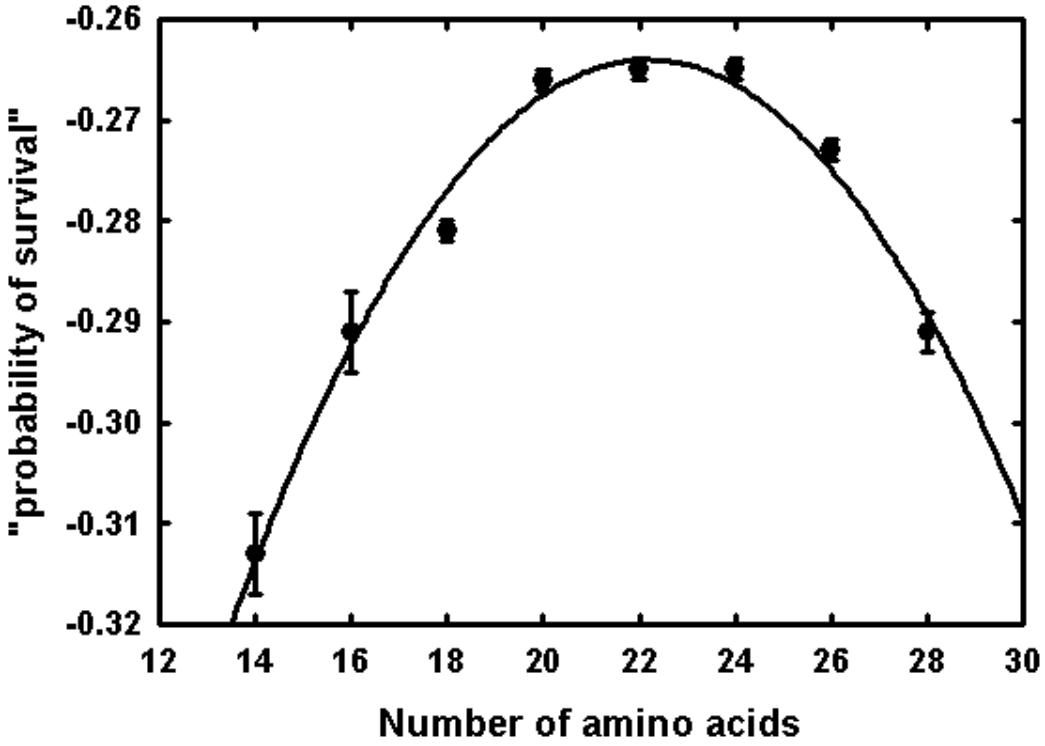

Fig. 3. "probabilities of survival" of different genetic code tables. The fit to a parabola is just to giude the eye.

the Scripps Research Institute [15] and provides computational justification for their belief that genetic codes with even more amino acids is feasible. However the number of amino acids will be restricted to 20-22-24 if we want the resulting life form to be resilient against mutations.

6 ACKNOWLEDGEMENTS

I am grateful to Dr. Isil Aksan Kurnaz and Dr. Muhittin Mungan for their contributions on the model and the calculations.

References

[1] T. J. P. Penna, J. Stat. Phys. **78**, 1629 (1995).

[2] D. Stauffer, Physica A **273**, 132 (1999).

[3] D. Stauffer, P. M. C. de Oliveira, S. M. de Oliveira and R. M. Z. dos Santos, Physica A **231**, 504 (1996).

[4] D. Stauffer, P. M. C. de Oliveira, S. M. de Oliveira, T. J. P. Penna and J. S. Sa Martins, Anais Da Academia Brasileira De Ciencias **73**, 15 (2001).

[5] A. O. Sousa, S. M. de Oliveira and J. S. Sa Martins, Phys. Rev. E **67**, 32903 (2003).

[6] E. Tuzel, V. Sevim and A. Erzan, Proc. Natl. Acad. Sci. **98**, 13774 (2001).

[7] E. Tuzel, V. Sevim and A. Erzan, Phys. Rev. E **64**, 061908 (2001).

[8] B. Orcal, E. Tuzel, V. Sevim, N. Jan and A. Erzan, Int. J. Mod. Phys. C **11**, 973 (2000).

[9] T. J. P. Penna, A. Racco, A. O. Sousa, Physica A **295**, 31 (2001).

[10] T. J. P. Penna and D. Stauffer, Zeitschrift Fur Physik B-Condensed Matter **101**, 469 (1996).

[11] T. J. P. Penna, S. M. de Oliveira and D. Stauffer, Phys. Rev. E **52**, R3309 (1995).

[12] Z. F. Huang and D. Stauffer, Theory in Biosciences **120**, 21 (2001).

[13] http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=nucleotide&val=285912.

[14] T. H. Jukes and C. R. Cantor *in Mammalian Protein Metabolism, edited by H. N. Munro* (Academic Press, New York, 21 1969).

[15] J. C. Anderson, N. Wu, S. W. Santoro, V. Lakshman, D. S. King and P. G. Schultz, Proc. Natl. Acad. Sci. **101**, 7566 (2004).

[16] T. J. Matthews, *Biochemistry* (Addison Wesley Longman, San Fransisco, 2000).

[17] J. Wang and W. Wang Nature Structral Biology **6**, 1033 (1999).

[18] S. Miyazawa and L. R. Jernigan J. Mol. Biol. **256**, 623 (1996).

[19] L. R. Murphy, A. Wallqvist and R. M. Levy, Prot. Eng. **13**, 149 (2000).

[20] S. Henikoff and J. G. Henikoff Proc. Natl. Acad. Sci. **89**, 10915 (1992).